

Sub-Nyquist Sampling and Moire-like Waveform Distortions

Glenn L. Williams

National Aeronautics and Space Administration

John H. Glenn Research Center at Lewis Field

Cleveland, OH 44135

ABSTRACT

Investigations of aliasing effects in digital waveform sampling have revealed the existence of a mathematical field and a pseudo-alias domain lying to the left of a "Nyquist line" in a plane defining the boundary between two domains of sampling. To the right of the line lies the classic alias domain. For signals band-limited below the Nyquist limit, displayed output may show a false modulation envelope. The effect occurs whenever the sample rate and the signal frequency are related by ratios of mutually prime integers. Belying the principal of a 10:1 sampling ratio being "good enough", this distortion easily occurs in graphed one-dimensional waveforms and two-dimensional images and occurs daily on television.

1. INTRODUCTION

As in most analog systems, high frequency roll-off effects in analog strip-chart recorders are seen in sine wave recordings when the signal frequency is increased beyond the electronic and electro-mechanical limits of the recorder. This usually occurs after the envelope of the recorded sine wave merges into a solid stripe. Modern digital waveform recorders and oscilloscopes contain very fast analog-to-digital converters in the signal-processing portion, so that no longer is the analog low-pass frequency roll-off of any concern. Waveform display or playback is nearly universally accomplished by graphing vector or raster line approximations of the recorded waveform on the display or chart. But for analog playback, the data may still be sent through a digital-to-analog converter, followed by resistive-capacitive (RC) or $\sin(x)/x$ filtering to smooth the analog output.

The Shannon (WKS) Sampling Theorem has been cited in countless articles as justification for dropping concern about sampling distortion effects. As long as the signal frequency bandwidth is well below the Nyquist frequency, common phrasing would say the "waveform can always be reconstructed completely because of Shannon's Theorem."

We will soon learn that neglecting the details of sampling theory can lead to serious misinterpretations regarding the display of certain waveforms under special conditions. Even with pure band-limited sine wave input, serious envelope distortions can occur. The distortions occur because most cost-effective commercial signal waveform playback and display products leave out the complete waveform reconstruction required in accordance with the Sampling Integral[3]. As a result, when technical accuracy is of strategic importance, particular care must be applied to interpreting the display and reproduction of sampled-data waveforms on oscilloscopes, waveform recorders, spreadsheet charts and even television.

This is a preprint or reprint of a paper intended for presentation at a conference. Because changes may be made before formal publication, this is made available with the understanding that it will not be cited or reproduced without the permission of the author.

2. THEORY

Adequate reconstruction of an unaliased waveform involves coping with mathematical and physical nuances, such as:

- (a) the signal must be continuous forever in the past and in the future, in order to obtain all the sample points necessary to perform the reconstruction, and
- (b) an infinitely fast digital processor is required to make all the computations in anything near real-time.
- (c) The Gibbs phenomenon[7] interferes with the final outcome.

Given these requirements a user can hardly be blamed for ignoring reconstruction issues when using data from a sampling system.

2.1 DEFINITION

For this discussion, a *distortionless* representation of a signal on an oscilloscope screen or waveform recorder output means that, except for a constant scaling factor, the amplitude of the displayed waveform envelope has the same "shape" as if the original signal were sampled at an infinite frequency.

2.2 DISCUSSION OF PLAYBACK DISTORTION

Digital recording instrument manufacturers commonly list in their sales literature a "flat" frequency response for their product. This claim is made due to the implicitly fast digital logic and the very high sample rate used in the recorders. The claim of "flat response" is, again, a result of the common misunderstanding regarding the Shannon Sampling Theorem. Both the sales engineer and the customer may believe that a properly bandwidth-limited digitally sampled signal can be reconstructed almost perfectly, right up to the Nyquist limit, and the instrument does reconstruction adequately. For some uses, textbook discussions[4] show that adequate reconstruction of the original waveform by analog means is theoretically possible and intuitively a requirement. However, the inherently high design and fabrication costs to include wide-band adaptive analog antialias filters can increase bottom-line price increase and discourage customers. And, it can be counterintuitive to have an analog output device in a digital waveform display system.

3. UNCOVERING A NEW DISTORTION

A recorder purchased and advertised as having a "flat" frequency response was set up to plot a wide frequency sine wave input sweep generating a nearly solid waveform stripe which would document the "flat response of the system". The signal source was from a high-quality sine wave generator from a different manufacturer. Soon a "beat" in the waveform display was observed near a certain single frequency on the signal generator (Figure 1). After "tuning" the frequency to reduce the frequency of the beat, and the chart printing drive was enabled. The obvious depth of the notch in the distorted envelope became an immediate concern since both the signal generator and the strip-chart recorder to be in otherwise excellent operating condition.

Since the Nyquist limit was not being exceeded, presumably no aliased waveforms were being plotted. The observed resonance effect and the envelope distortion were therefore a bit disconcerting. Further experimentation showed that the effect was not a fluke. A whole family of frequencies showing this effect were recorded and logged. This test was repeated with different devices, using a digital storage oscilloscope made by an unrelated manufacturer, and input from a different sine wave generator also made by a different unrelated manufacturer. The results were consistently repeatable.

Later study of the frequencies involved resulted in the development of a mathematical model explaining what had happened. Various texts and papers on sampling theory, Moire' patterns, and the like were researched, without ever locating a similar underlying model. In this model, a representative amount of envelope distortion occurs at a signal frequency which is 6/53 (0.113207...) of the sample rate, or roughly 1/9 of the sample rate. This result is *always reproducible* with recording equipment which is in good operating condition. This result is also reproducible on a computer spreadsheet. The reader is invited to perform an independent trial of this experiment.

3.1 INTRODUCTION TO THE MODEL

A constant frequency and constant amplitude sine wave waveform is most certainly a bandwidth-limited waveform, expressed as

$$f(t) = \sin(\omega t) \quad (1)$$

where

$$\omega = 2\pi f$$

Assume that this waveform is sampled such that

$$f \leq f_c = \frac{1}{2} f_s \quad (2)$$

where f_c is the Nyquist frequency and f_s is the sample rate. Then we know that the Nyquist limit is not being exceeded. Now we simplify Equation 2 as

$$f \leq \frac{1}{2} f_s \quad (3)$$

We then generalize by defining arbitrary positive real integers m and n such that

$$f \leq \frac{m}{n} f_s \quad (4)$$

with the constraint that

$$m \leq \frac{1}{2} n \quad (5)$$

or, as a redefinition of the Nyquist limit,

$$2m \leq n \quad (6)$$

The initial experimental choice of 6 and 53 for m and n in Equation 4 happen to yield a large, obvious distortion. The reasoning applies to the general case of any two mutually prime integer numerators and denominators constrained as in Equations 4 or 5.

Of course, if m is exactly one-half of n , the Nyquist frequency is being sampled and the output is useless. For $\frac{m}{n} = 1/3, 1/4$, etc. similar ugly results are obtained. But there are dozens of other highly visual possibilities in the range where

$$\frac{1}{20} < \frac{m}{n} < \frac{1}{2} .$$

The reader should be aware that this problem occurs in all sampling situations where there is a constant frequency sample rate or a constant spatial distance between samples. Other fields in which sampling of variable data occurs are numerous, such as meteorology, medical research, economics, statistics. For instance, assume a key economic indicator is analyzed monthly by an economist, and the results are published in the economics literature. The economist maintains that the indicator has a hitherto unrecognized cycle, a cycle which is almost 106 months long. But 106 months of a monthly published statistic could be related by 106/12 or 53/6. An obvious distortion in the data could occur having nothing at all to do with a real cycle or a real fact.

3.2 VIDEO (TELEVISION) EXAMPLES

The pixel size in a video display partially defines the *spatial* sampling system having a *spatial frequency*. The analog signals derived from the demodulation of the radio-frequency television carrier signal add a time element to the display. Sampling effects in the video image also occur because of the vertically interleaved raster lines, the field and framing rates, and the vertical retrace intervals, all of which form other time and spatial frequencies for sampling.

A common sampling problem begins when the video image happens to contain a fine-grained repetitive pattern, so that the spatial pitch (frequency) of the image pattern will be exactly a fractional portion of the cathode ray tube shadow mask sampling frequency, or the vertical raster pitch. Repetitive patterns occur in all sorts of ways, such as views of venetian blinds, clothing with patterns, the flag of the United States, etc. Figure 3 shows an example of such an image having a large area with a rapid spatial frequency and area of confusion. The image is that of a person wearing a jacket with a tight hound's tooth or herringbone twill weave. Most people have seen these effects on television. Such effects can be found daily, and the effects can be quite annoying. Television hardware manufacturers often claim their hardware minimizes the effects of these Moire' patterns.

In Figure 4, a 512x484 virtual image was synthetically created with a C program. Each of the 484 raster lines of the image are identical, and result from calculating and repeating a raster line containing 512 samples of a sine wave waveform having a frequency which is 6/53 of the sample rate. A few extra dark vertical bands corresponding to the distortion notch are visible as vertical stripe patterns.

4. MORE RESULTS FROM THE MODEL

4.1 PERCENTAGE OF DISTORTION ERROR

The "rule of thumb" often used to avoid severe sampling distortion near the Nyquist limit is to have a 10:1 ratio, i.e.

$$\frac{n}{m} \geq 10 .$$

We now show that the 10:1 rule of thumb is not very practical if the system performance specification requires that no detectable distortion. Due to the large numbers possibilities which could yield significant "envelope modulation", it is necessary to understand how to calculate the amplitude ripple (notch depth of the cusps).

The percentage of modulation is derivable as follows from the maximum phase difference

$$\theta = \cos\left(\frac{\pi m}{n}\right). \quad (7)$$

From the graphed figures, it is clear that the overall period of the modulation error is $m * n$. It is also clear that the time $m * n$ represents the entire cycle of the modulation error, and therefore, is the time or horizontal distance from one peak to the next peak of the modulation ripple. The number of samples from peak to valley of the modulation is then equal to

$$\frac{m * n}{2}. \quad (8)$$

Starting from time zero, after approximately one full cycle of the signal at rate m , exactly n samples will have been acquired.

We now convert from time units to radians of revolution. A full cycle of the signal is 2π radians, and each sample will be spaced every $\frac{2\pi}{n}$ radians. At the end of n samples, there will be a residual angle left over on the last sample. The residual must applied as a debt or credit to the following signal cycle of 2π radians. This residual slowly accumulates over many periods to reach a maximum where the valley of the ripple has the greatest depth. Therefore, the total length from peak to valley is, converted from above,

$$(\text{one-half \# of samples}) * (\text{distance between samples}) = \left(\frac{m * n}{2}\right) * \left(\frac{2\pi}{n}\right) \text{ radians.} \quad (9)$$

There is a temptation to simplify Equation 9, but let's not do it yet. We need to find the left over phase error at the valley, which is calculated by calculating the residual phase error modulo m :

$$\theta = \left[\left(\frac{m * n}{2} \right) \left(\frac{2\pi}{n} \right) \right] \text{mod}(m) = \left(\frac{2\pi}{n} \right) \left[\left(\frac{m * n}{2} \right) \text{mod}(m) \right] \quad (10)$$

Now, we need to make use of the fact that m and n are mutually prime integers. Since, per Equation 6, m is less than n , then m is the only one of the two integers which can take on the value 2. Therefore, n , being greater than two and prime, will always be odd, and $n/2$ is also not a whole number. But, if n is odd, then $(n - 1)$ is even, and $(n - 1)/2$ is a whole number. Then,

$$\left(\frac{m * n}{2} \right) \text{mod}(m) = \left[(n - 1 + 1) * \frac{m}{2} \right] \text{mod}(m) = \left[\left((n - 1) * \frac{m}{2} \right) + \left(\frac{m}{2} \right) \right] \text{mod}(m) \quad (11)$$

$$\begin{aligned} &= \left[m \left(\frac{n - 1}{2} \right) \right] \text{mod}(m) + \left(\frac{m}{2} \right) \text{mod}(m) \\ &= 0 + \frac{m}{2} \end{aligned} \quad (12)$$

and therefore, the total phase error is

$$\theta = \left(\frac{m}{2} \right) \left(\frac{2\pi}{n} \right) = \pi \left(\frac{m}{n} \right). \quad (13)$$

The normalized full-scale amplitude of the signal in the valley of the notch, α_v , is then

$$a_v = \cos(\theta) = \cos\left(\frac{\pi m}{n}\right) \quad (14)$$

Using the previous example, $\frac{m}{n} = 6/53$, and $a_v = 0.937...$ which amounts to a 6% peak error!

Given that amount of error amplitude, we are curious to evaluate when the distortion error is smaller than a certain percentage of full scale, more importantly, when the error is less than the 1-bit quantization size in an n -bit sampling system. Referring to Figure 1, for the normalized full-scale error value ε such that:

$$\varepsilon = 1 - \alpha_v = 1 - \cos\left(\pi * \frac{m}{n}\right) \quad (15)$$

when applied to an 8 bit sampling system, gives an error ε smaller than one bit which is

$$1/256 = 0.4\%.$$

Solving for the ratio m/n in Equation 15

$$\frac{m}{n} < \frac{1}{\pi} * \arccos(1 - \epsilon) \quad (16)$$

and applying Equation 16 to the above error of 0.4%,

$$\frac{m}{n} < 0.028.$$

Therefore, for a telephone line channel sampled at 8000 times per second (a common telephone industry sample rate for a private line) with an 8-bit analog-to-digital converter, distortion is negligible below 224 Hz, a frequency lower than the musical note middle C.

Similarly, for a 10 bit waveform recorder (1000 points across the print head) sampling at 10,000 Hz, a manufacturer might claim that the instrument is "flat to 5000 Hz," which is the Nyquist frequency. So to display no modulation above one dot peak error (.1%) at full scale requires having no signals exceeding 142 Hz! Typically, instrument manufacturers will be reluctant to admit this constraint exists because, through no fault of their own, this constraint would make the specifications for their instrumentation equipment appear to be rather poor!

4.2 DISCUSSION OF THE NYQUIST THEOREM

One might ask an almost obvious question at this point. Why not just "correct" the distortion effects by putting in a variable amplification factor to remove the dip in the waveform? The answer to that is two-fold. First, there is no *a priori* way to know the locations of the distortions when in most situations the signal frequencies are random. Second, from Fourier theory, we know that a typical waveform is not a perfect sine wave, but is really a composite of many waveforms, and may contain some level of random noise from the "real world". Therefore the offending distortion, and only the offending distortion, would have to be isolated first in order to perform the correction. It would be far easier to implement the reconstruction via the Sampling Integral and let all the corrections be done with a consistent provable algorithm. (There is still the noise problem.)

Our past attempts to perform waveform reconstruction on a computer, given unlimited computation time, and any number of computer languages, has shown no useful solution to the reconstruction problem. Given a finite sequence of samples, we have been able to apparently reconstruct a sine wave without modulation distortion. But the very same algorithms fail miserably to reconstruct square waves, and we can infer that other waveforms such as sawtooth waveforms would also show severe reconstruction and Gibbs phenomenon errors.

4.3 GRAPHING THE FIELD OF POSSIBILITIES

The ratios m/n form a mathematical field of fractions where the numerators and denominators are mutually prime integers or products of mutually prime integers. When graphed in Cartesian coordinates such that n is the ordinate and m is the abscissa, and little crosses (\times), or *nodes*, mark each $\{m, n\}$ coordinate, the resulting graph resembles a minefield of nodes (Figure 5). The nodes representing the Nyquist frequency can be connected as a line of slope equal to 2 crossing the origin at

$\{0,0\}$, hereby named the *Nyquist line*. For clarity, Figure 5 has been drawn so that nodes on the right-hand side of the Nyquist line ($m > n/2$) have been excluded.

In Figure 5, the regions on each side of the Nyquist line have been labeled as "I" and "II" for a particular reason. Region I is the region where classic signal aliasing is defined. Region II defines the domain where sub-Nyquist distortions, the subject of this paper, are defined. It is in this region where useful waveform display occurs. The ratio between a signal frequency and the sampling frequency is for all practical purposes a "random" ratio somewhere in Region II, and hopefully misses all the nodes. In Fourier theory, non-sinusoidal signals contain several or many sinusoidal components. One or more of the component harmonic frequencies of a useful waveform could happen to lie very near a Region II node. It would then be possible for those components to be severely distorted. The importance and seriousness of the modulation would depend on the amplitude of the affected component relative to the overall signal amplitude.

A conjecture will now be made. It is also possible to consider that the envelope distortions which occur in Region II are really a form of alias, a "sub-Nyquist" alias. Therefore, in this report, Region I aliases could be named "Type I aliases." Similarly, Region II modulations could be named "Type II aliases."

5. CONCLUSIONS

In waveforms, statistical data, and images, sample aliasing due to system coincidences can occur in one of two domains as defined by the Nyquist frequency line in Figure 5. The two domains have the characteristic that:

- a) Classical aliasing in Region I occurs anytime where m and n are defined as in Equation 4, i.e., m and n are real numbers such that $n \geq 2m$.
- b) Region II distortion occurs for any positive real integers m and n where

$$1 \leq m \leq \frac{n}{2} \quad \text{where}$$

$n \geq 2$ and m and n are mutually prime. The integers m and n separately can be products of prime numbers without affecting these conclusions. These mutually prime integers and products m and n form a mathematical field of possible trouble points, or nodes. System designers should avoid Region II nodes whenever possible if they are implementing systems which do not perform adequate waveform reconstruction before presenting plots which display critical information.

This report has shown that a new domain of distortion, in Region II, has subtle implications for the fabrication of systems using digital waveform sampling. Except for television, where the effects of swimming color bands are obvious and even obnoxious, there has not been a great deal of attention focused on this type of alias. However, in the future, engineers and statisticians should determine what impact the Region II distortion may have their data before drawing conclusions.

Finally, in this report, no detailed analysis has been done to see if the modulation effects around a Region II node result in extra peaks in the power spectrum indicating signal power is aliased into undesirable frequencies. No claim is made that the Region II distortions result in real signal power

being lost from the sampled signal. However, we are concerned that in rare cases the envelope distortions could be interpreted as modulation and cause serious consequences in error detection systems and feedback control systems.

6. REFERENCES

1. C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE*, Vol. 37, pp. 10-21, Jan. 1949.
2. Ahmed I. Zayed, "Advances in Shannon's Sampling Theory", CRC Press, New York, 1993.
3. Chi-Tsong Chen, "One-Dimensional Digital Signal Processing," Marcel Dekker, Inc., 1979, pp. 69-83.
4. Mischa Schwartz, "Information, Transmission, Modulation, and Noise", McGraw-Hill, 1959, Chapter 4.
5. "Data Acquisition and Conversion Handbook", edited by Eugene L. Zuch, ca. 1978, published by Datel-Intersil Corp., p. 236.
6. Abdul J. Jerri, "The Shannon Sampling Theorem -- Its Various Extensions and Applications: A Tutorial Review", *Proc. IEEE*, Vol. 65, No. 11, Nov. 1977, pp. 1565-1596.
7. D. Gottlieb, C.-W. Shu, A. Solomonoff and H. Vandeven, "On the Gibbs Phenomenon I: recovering exponential accuracy from the Fourier partial sum of a nonperiodic analytic function", *J. Comput. Appl. Math.*, v43, 1992, pp. 81-92. (See also subsequent papers II, III, and IV.).
8. P. Mertz and F. Gray, "A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television", *The Bell System Technical Journal*, Vol. 13, July 1934, pp. 464-515 (in "Graphical and Binary Image Processing and Applications", edited by J. C. Stoffel, Artech House, 1982, pp. 5-56.).

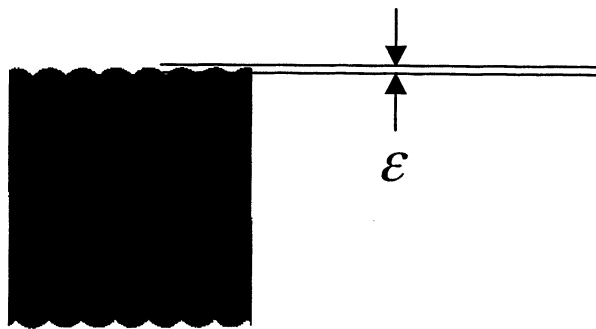


Figure 1. An actual waveform captured with a commercial digital thermal strip chart recorder. The sample rate in the strip chart recorder was measured as 12170 Hz. The signal was from a sine-wave signal generator set to 1367 Hz. Thus the ratio of signal frequency to sample rate was 0.1130649, or almost exactly 6/53. The maximum percentage modulation error ϵ depends on determining the distribution of samples near the envelope peak, as in Equation 15.

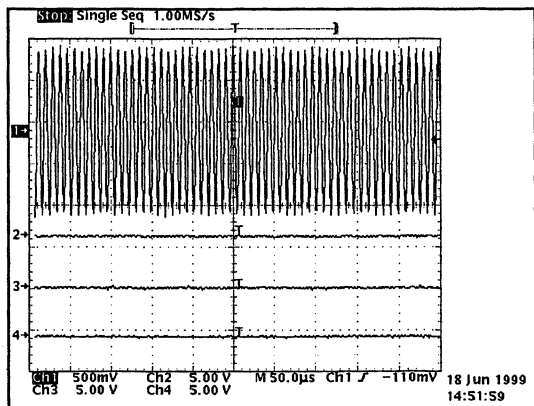


Figure 2. An waveform captured with a commercial digital storage oscilloscope and saved as a graphics file. The sample rate is shown. The sine wave input frequency on Channel 1 was 113.20754 kHz from a commercial 1 Volt peak-to-peak synthesized waveform generator.

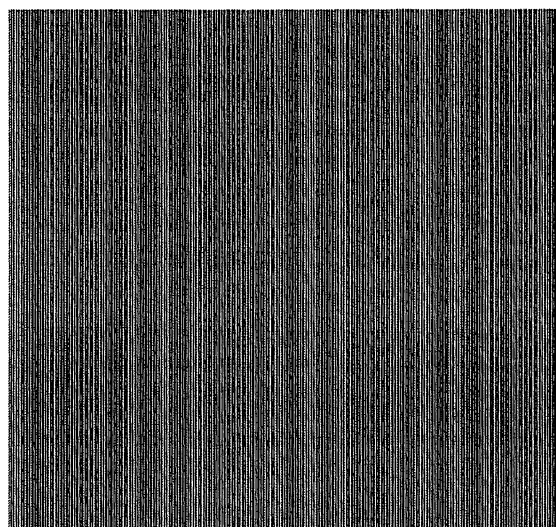


Figure 4. A 512 x 484 video image is carefully synthesized from a raster line which is a raised sine wave sampled in the ratio 6/53. NOTE: The exact video effects are often obscured by Moiré effects induced by the software and hardware used to actually print this image. In addition, this image has been given a histogram stretch to emphasize the banding.

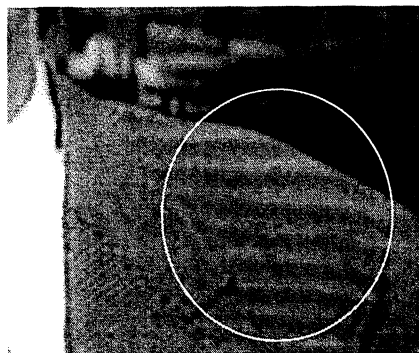


Figure 3. This image was photographed off the screen of a color television with a 35mm film camera. An 8x10 print of the negative was scanned on a high resolution scanner and cropped. It should be noted that this figure is neither a proof that the patterns are caused only by the Moiré effect, nor that they are specifically caused by sub-Nyquist distortions.

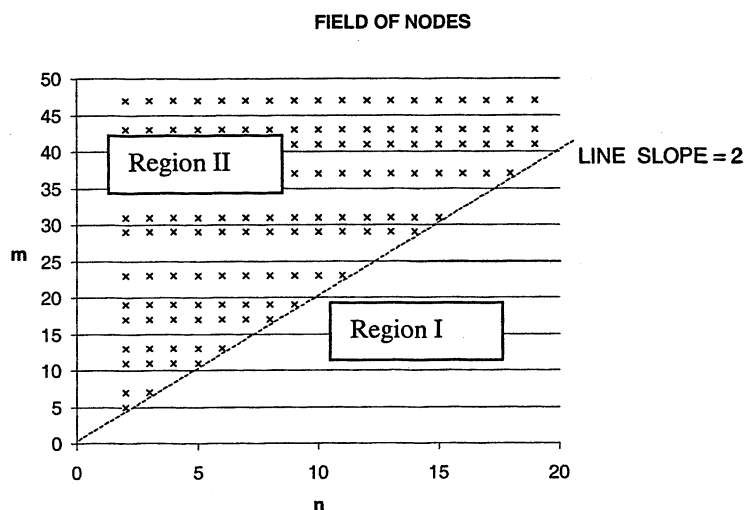


Figure 5. A subset of the infinite field of points (nodes)

representing the prime fraction $\frac{m}{n}$ is shown above. Both the domain and range of the set of points extends to infinity, but the numerator m becomes more sparse. This field shows possibilities such as $\frac{53}{6}$ which is of type $\frac{m}{n_1 * n_2}$ but does not show equally valid possibilities like $\frac{m_1 * m_2}{n_1 * n_2}$ and so forth.